МАШИННЫЕ МЕТОДЫ ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
1981 год Выпуск 88

УДК 577.2

МЕТОД РАСЧЕТА НИЗКОЭНЕРГЕТИЧЕСКИХ ВТОРИЧНЫХ СТРУКТУР РНК, ИСПОЛЬЗУЮЩИЙ АЛГОРИТМ ПОИСКА КЛИК

Л.В.Омельянчук, Ю.Е.Бессонов, Н.А.Колчанов

Рибонуклеиновые кислоты (РНК) - биополимеры,состоящие из нуклеотидов четырех типов: аденина, урацила, гуанина и цитозина. Молекула РНК является линейной, т.е. нуклеотиды соединены между собой последовательно, образуя цепочку. Пары "аденин - урацил" или "гуанин - цитозин" называются комплементарными, так как образующие из нуклеотиды, сближаясь в пространстве, могут формировать водородные связи. Два отрезка полинуклеотидной цепи, сближаясь в пространстве при изгибании молекулы РНК, образуют непрерывную последовательность комплементарных пар и формируют двойную спираль Уотсона-Крика [I]. Неспиральные участки молекулы РНК образуют петли. Вторичная структура РНК представляет собой набор двойных спиралей и петель.

Известно [I,2], что вторичная структура РНК играет важную роль в функционировании молекулярно-генетических систем управления, и в то же время экспериментальные методы ее определения сложны и трудоемки. Единственный пример [I] прямого экспериментального определения вторичной структуры РНК - это рентгено-структурный анализ т-РНК^{фен} из дрожжей. Во всех остальных случаях экспериментальная информация является косвенной и может быть получена либо по данным биохимического анализа РНК,либо путем измерения некоторых физических характеристик РНК в растворах.

Особый интерес представляет выявление наборов наиболее низкоэнергетических конформаций вторичной структуры молекул РНК,которые обеспечивают биологическую активность этого класса макромолекул.Цель настоящей работы состояла в создании метода расчета наборов низкоэнергетических вторичных структур молекул РНК по задан-

ным полинуклеотидным последовательностям с учетом стандартных термодинамических характеристик формирования вторичных структур РНК [3,4].

Рассматриваемый метод отличается от предложенных ранее [4] прежде всего тем, что перебор вариантов вторичной структуры осуществляется оптимальным способом с использованием алгоритма поиска клик [5].

Так как даже для сравнительно коротких последовательностей РНК общее число возможных вторичных структур крайне велико*), то для нахождения вторичной структуры РНК, соответствующей минимуму свободной энергии, необходимо построение подходящей модели процесса формирования вторичной структуры РНК и использование алгоритмов, отличающихся от полного перебора.

§I. Основные понятия и обозначения

Представим молекулу РНК упорядоченной последовательностью $B = (a_1, a_2, \ldots, a_n)$ символов $a_i \in \{A, У, Г, Ц\}$. Пару (a_i, a_j) назовем комплементарной, если $\{a_i, a_j\} = \{A, У\}$ или $\{a_i, a_j\} = \{Г, Ц\}$, где $i \in [1,n]$, $j \in [1,n]$ и $i < j - 3$.

Смысл последнего ограничения состоит в том, что нуклеотиды, расположенные в цепи близко друг от друга, не могут образовывать водородные связи.

ОПРЕДЕЛЕНИЕ I. Двойной спиралью будем называть максимальный по включению набор комплементарных пар

$$S = ((a_{i_1}, a_{j_1}), (a_{i_2}, a_{j_2}), \ldots, (a_{i_k}, a_{j_k})),$$

удовлетворяющий условиям: $i_l = i_1 + l - 1$, $j_l = j_k + k - l$, где $1 \le l \le k$ и $k > 1$.

Две упорядоченные последовательности $B' = (a_{i_1}, \ldots, a_{i_k})$ и $B'' = (a_{j_k}, a_{j_{k-1}}, \ldots, a_{j_1})$, формирующие двойную спираль, будем называть взаимно комплементарными.

Существуют два ограничения на одновременное присутствие двух произвольных двойных спиралей в одной вторичной структуре. Эти ограничения отражают стереохимические особенности вторичных структур [2] и схематически представлены на рис.I. Нуклеотиды обозначены точками. Соседние по цепи нуклеотиды соединены непрерывной

*) Для молекулы РНК, содержащей 80 нуклеотидов, общее число возможных вариантов образования наборов комплементарных пар может составлять 10^{30}.

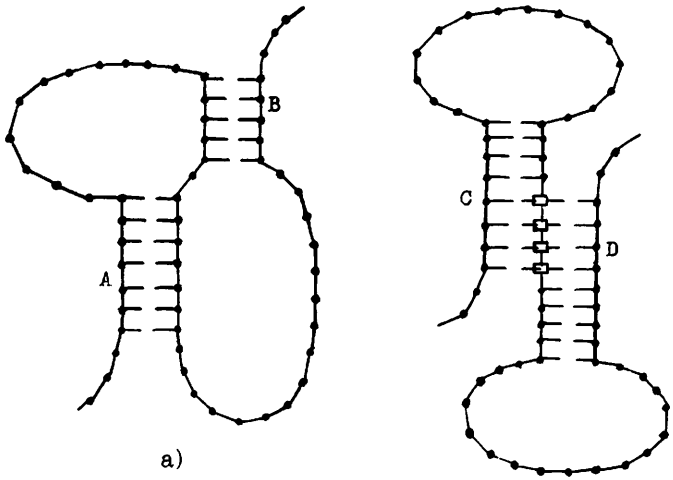

Рис. I

линией. Комплементарные нуклеотиды, образующие спирали A,B,C,Д, соединены пунктиром. а) Спирали А и В взаимно запрещены, так как один из комплементарных участков спирали А расположен в цепи между двумя комплементарными участками спирали В,и наоборот; б) Спирали С и Д взаимно запрещены, так как имеют нуклеотиды, образующие по две комплементарных пары (обозначены квадратами).Используя введенные выше обозначения, эти ограничения можно сформулировать следующим образом:

ОПРЕДЕЛЕНИЕ 2. Две спирали S_1 и S_2 взаимно разрешены, если для каждой пары $(a_i,a_j) \in S_1$ и каждой пары $(a_r,a_s) \in S_2$ имеет место

$$(r < i) \& (j < s) \vee (i < r) \& (s < j) \vee (j < r) \vee (s < i). \qquad (1)$$

Обозначим множество всех возможных двойных спиралей полинуклеотидной последовательности В через W(B) . Отношения стереохимической разрешенности можно представить графом $H_B = (V,E)$,в котором мнюжество вершин V представляет спирали, а множество ребер E - взаимно разрешенные пары спиралей.

ОПРЕДЕЛЕНИЕ 3. Любой набор взаимно разрешенных спиралей называется вторичной структурой.

Очевидно, каждой вторичной структуре последовательности B будет соответствовать некоторый полный подграф графа $\bar{n}_B$.

Пусть вторичная структура задана набором взаимно разрешенных спиралей $\Omega = \{S_k\}_{k=1,\ldots,M}$, причем каждая спираль S_k задана двумя комплементарными участками B'_k и B''_k. Определим типы неспиральных участков, которые могут возникнуть во вторичной структуре. Для этого построим граф T по следующим правилам:

1) каждому нуклеотиду РНК поставим в соответствие вершину графа T;

2) две вершины соединим ребром типа I, если два соответствующих нуклеотида являются соседними в цепи;

3) две вершины соединим ребром типа П, если два соответствующих нуклеотида являются комплементарной парой некоторой спирали S_k.

ОПРЕДЕЛЕНИЕ 4. Петлей вторичной структуры назовем любой цикл графа T, который содержит не более одной вершины, принадлежащей одному комплементарному участку некоторой спирали. Длину $|C|$ петли C определим как число вершин цикла, не входящих в комплементарные пары спиралей.

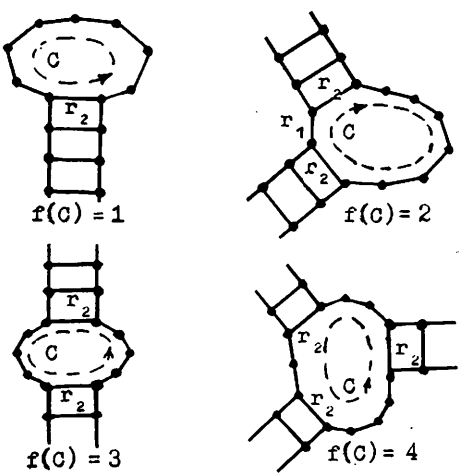

Рис.2

Возможны следующие типы петель:

I) Цикл C содержит одно ребро типа П. Тогда соответствующую петлю назовем шпилечной и присвоим ей индекс $f(C) = 1$.

2) Цикл C содержит два ребра типа П и цепь $r_2 r_1 r_2$, где r_1 - ребро типа I и r_2 - ребро типа П. Петлю назовем боковой и положим $f(C) = 2$.

3) Цикл C содержит два ребра типа П, но не содержит цепи вида $r_2 r_1 r_2$. Назовем петлю внутренней и положим $f(C) = 3$.

4) Цикл C содержит более двух ребер типа П. В этом случае назовем петлю с л о ж н о й и положим f(C) = 4.

На рис. 2 изображены все типы петель.

Каждая вторичная структура может быть охарактеризована определенным значением свободной энергии, которое вычисляется как сумма отрицательной свободной энергии двойных спиралей и положительной свободной энергии петель.

Энергия двойной спирали S_k есть

$$G(S_k) = \sum_{p=1}^{m-1} \Delta G(a_{i_p}, a_{j_p}, a_{i_{p+1}}, a_{j_{p+1}}), \qquad (2)$$

где $\Delta G(a_{i_p}, a_{j_p}, a_{i_{p+1}}, a_{j_{p+1}})$ - энергия взаимодействия двух соседних комплементарных пар, входящих в двойную спираль, m - длина спирали. Значения этих энергий при всех возможных ориентациях комплементарных пар определены экспериментально.

Энергия петли C зависит от ее типа f(C) и длины $|C|$. Значения энергий для всех типов петель в зависимости от их длины определены экспериментально.

Таким образом, энергия вторичной структуры есть

$$G = \sum_{k=1}^{M} G(S_k) + \sum_{C} G'(|C|, f(C)), \qquad (3)$$

где $G'(|C|, f(C))$ - энергия петли типа f(C) и длины $|C|$. Поскольку каждой вторичной структуре соответствует некоторый полный подграф в $H_B(W)$, то в дальнейшем под энергией полного подграфа будем понимать энергию соответствующей ему вторичной структуры.

§2. Метод расчета низкоэнергетических вторичных структур

Непосредственным перебором всех возможных вариантов типов и длин петель было установлено существование энергетического порога U = -6 ккал/моль, позволяющего выделить из множества W всех возможных для данной последовательности B спиралей подмножество высокостабильных спиралей $W^* = \{S \in W \mid G(S) \le U\}$, такое, что для двух наборов стереохимически разрешенных спиралей $\Omega, \Omega' \subset W^*$, удовлетворяющих условию $\Omega \subset \Omega'$, всегда имеет место:

$$G(\Omega) > G(\Omega'). \qquad (4)$$

Это свойство монотонности энергии обусловлено тем, что образование любой спирали из множества W^* является энергетически вы-

годным при любых типах и длинах петель, расположенных на ее концах.

Поскольку каждой вторичной структуре соответствует некоторый полный подграф графа $H_B(W)$, то свойство (4) означает, что энергия любого полного подграфа выше, чем энергия включающей его клики.

Наличие стереохимических ограничений (I) и свойство монотонности энергии (4) позволяет предложить следующую модель процесса формирования вторичной структуры молекулы РНК.

На первом этапе этого процесса формируется вторичная структура из высокостабильных спиралей. На втором этапе на неспиральных участках этой структуры формируются низкостабильные спирали, образуя энергетически наиболее выгодные вторичные структуры для каждого из этих участков.

Метод расчета низкоэнергетических конформаций вторичных структур РНК, по предложенной модели, состоит из двух этапов.

П е р в ы й э т а п.

I. Для заданной полинуклеотидной последовательности B определяется множество всех возможных спиралей W(B) и вычисляется энергия каждой спирали по формуле (2).

2. Выделяется множество высокостабильных спиралей $W^* \subset W(B)$, и строится граф взаимно разрешенных спиралей H_B^*, вершины которого представляют спирали из множества W^*.

3. В графе H_B^* отыскиваются все клики при помощи алгоритма [5]. Для каждой клики, соответствующей вторичной структуре, по формуле (3) вычисляется энергия и выбирается набор из n вторичных структур с минимальной энергией.*)

В т о р о й э т а п. Для каждой из n вторичных структур Ω_k ($k = \overline{1,n}$), полученных на первом этапе, выполняется следующее:

I. В структуре Ω_k выделяются все петли $B_1, B_2, \ldots, B_{p_k}$.

2. Для каждой последовательности B_i определяется множество низкостабильных спиралей $W(B_i) = \{S \mid G(S) > U\}$ и строится граф H_{B_i} взаимно разрешенных спиралей из $W(B_i)$.

3. Для каждого $i \in [1, p_k]$ в графе H_{B_i} отыскиваются все полные подграфы (не обязательно клики, так как для низкостабильной спирали свойство (4) не выполняется) и выбирается полный подграф, соответствующий структуре с минимальной энергией.

---

*) Величина n задается пользователем алгоритма. В приведенных ниже расчетах $n \sim 10$.

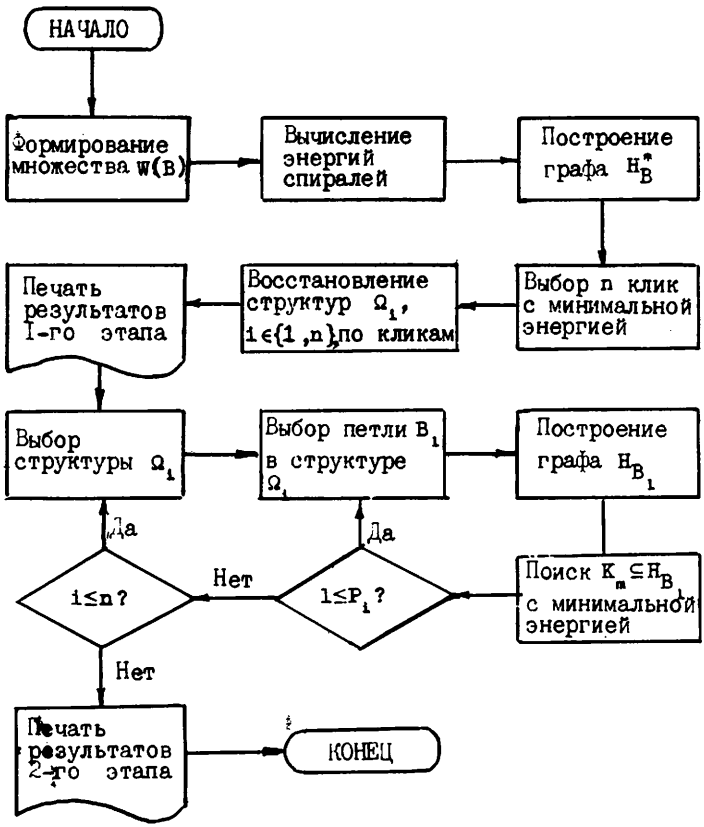

Рис.3. Блок-схема алгоритма расчета низкоэнергетических вторичных структур.

Алгоритм расчета вторичной структуры РНК, обладающий минимальной энергией, изображен в виде блок-схемы на рис.3.

Алгоритм реализован на ЭВМ БЭСМ-6 (язык АЛГОЛ ГДР). Среднее время обработки одной последовательности РНК, имеющей длину 80 нуклеотидов, составляет приблизительно 10 сек.

§3. Исследование вторичных структур природных РНК

Был проведен расчет 93-х т-РНК с известными полинуклеотидными последовательностями [6]. Приведем результаты расчета для т-РНК^{фен} из дрожжей. В данной нуклеотидной последовательности возможно формирование 120 спиралей. При пороге U = -6 ккал/моль остается лишь 19 низкоэнергетических спиралей. Граф H_B^* для этого

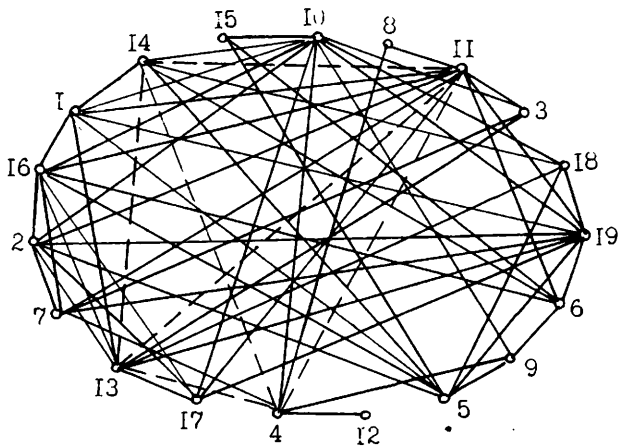

Рис. 4

набора спиралей изображен на рис.4. (На рис.4 вершины соответствуют низкоэнергетическим спиралям. Ребрами соединены взаимно разрешенные спирали. Пунктиром соединены спирали, входящие в максимальный полный подграф с минимальной свободной энергией, который соответствует структуре "клеверного листа".) Этот граф содержит 33 клики. Шесть наиболее низкоэнергетических структур, полученных на первом этапе минимизации, представлены на рис.5. Прямоугольниками обведены спирали, добавленные на втором этапе минимизации. Рассчитанная структура с наименьшей энергией совпадает со струк -

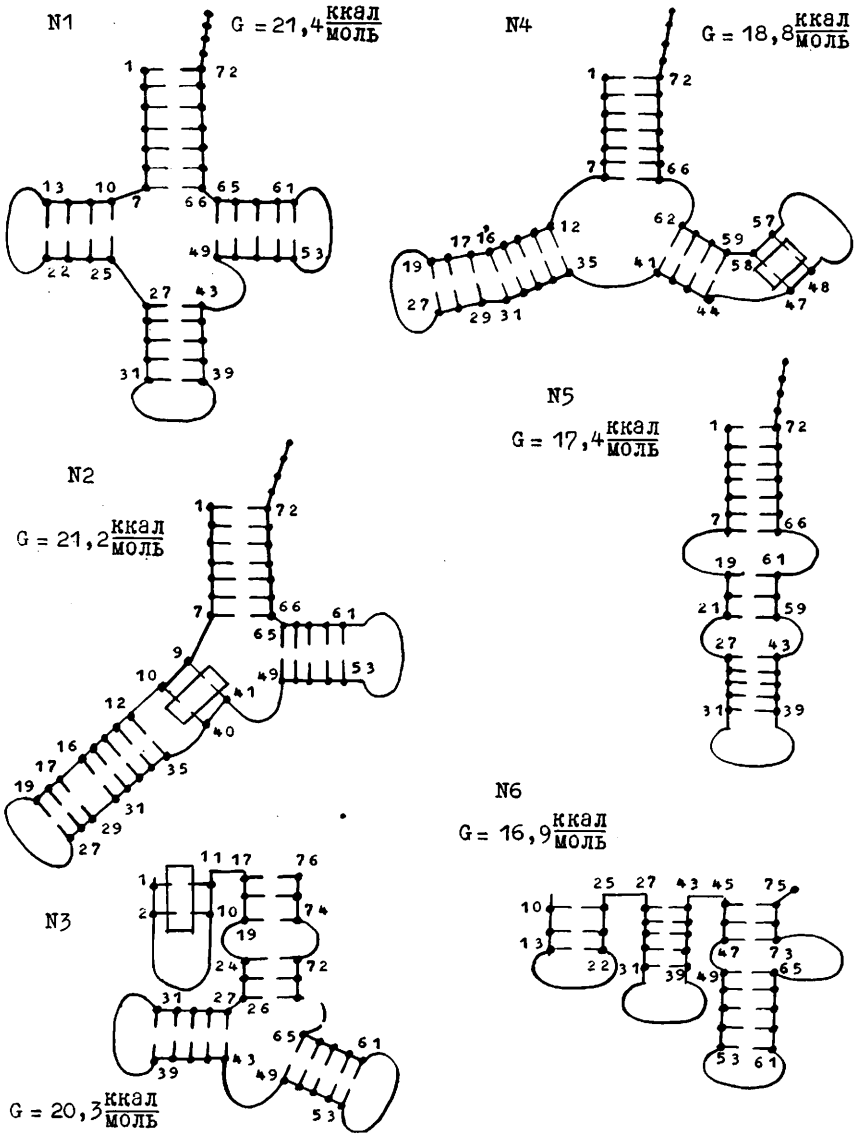

Рис.5. Шесть наиболее низкоэнергетических вторичных структур т-РНК^{фен}: G – величина свободной энергии; структура №I – "клеверный лист".

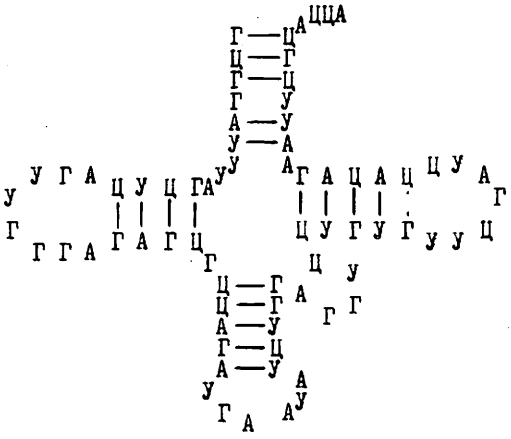

Рис.6

турой "клеверного листа", полученной из рентгеноструктурного анализа (рис.6).

Вторичная структура "клеверного листа" является инвариантной для всех т-РНК, так как в любой из них возможно одновременное формирование тех спиралей и петель, которые наблюдаются в т-РНК^{фен}.

Результаты расчета всех остальных т-РНК представлены в табл.I.

Здесь N - количество т-РНК, имеющих заданное отклонение ΔG свободной энергии "клеверного листа" от минимальной для каждой т-РНК. Для каждой т-РНК были рассчитаны наиболее низкоэнергетическая вторичная структура и разность между энергией этой структуры и энергией "клеверного листа". Совпадение структуры "клеверного листа" со структурой с минимальной энергией имело место для 72% проанализированных молекул т-РНК. Поэтому можно сделать вывод, что вторичная структура "клеверного листа" является одной из самых низкоэнергетических структур для большинства т-РНК.

Таблица I

ΔG ккал/моль	0	I-2	3-4	5-6	7-8	9-II	I2-I4
N	66	8	6	8	0	2	I

Была рассчитана вторичная структура 23-х рибосомных 5SРНК эукариот с известными полинуклеотидными последовательностями [7]. Для I6-ти 5SРНК существует инвариантная вторичная структура типа I, а для 7-ми типа II (рис.6). Для каждой 5SРНК были рассчитаны наиболее низкоэнергетическая вторичная структура и разность между ее энергией и энергией соответствующей инвариантной структуры (табл. 2). Здесь N - количество 5SРНК, имеющих заданное отклонение ΔG свободной энергии инвариантной вторичной структуры от минимальной для каждой 5SРНК.

Таблица 2

ΔG ккал/моль	4-7	7-I0	I0-I3	I3-I6	I6-I9	I9-25
N	3	9	7	4	I	I

Вторичные структуры, соответствующие минимуму свободной энергии, для всех 23-х 5SРНК, отличаются от инвариантных: каждая 5SРНК имела специфическую наиболее низкоэнергетическую вторичную структуру с уникальным расположением двойных спиралей и петель.

Несовпадение инвариантной вторичной структуры, которая, по-видимому, является функционирующей формой 5SРНК, со структурой с минимальной энергией обусловлено, возможно, тем, что в отличие от т-РНК (для которых такое соответствие имеет место) 5SРНК не в свободном состоянии функционирует, а лишь в комплексе с рибосомными белками. Можно думать, что 5SРНК образует инвариантную вторичную структуру именно в комплексе с рибосомными белками при формировании пространственной структуры рибосомы.

Анализ расчетов показывает (см. рис.5), что на втором этапе минимизации энергия вторичной структуры снижается незначительно. Это означает, что при дальнейших расчетах достаточно реализовать лишь первый этап минимизации.

Литература

I. УОТСОН Дж. Молекулярная биология гена. - М.: Мир, I978.- 525 с.

2. РАТНЕР В.А. Молекулярно-генетические системы управления.- Новосибирск: Наука, I975. - 273 с.

3. Improved Estimation of Secondary Structure in Ribonucleic Acids/ I.Tinoco, R.N.Borer, B.Dengler, M.D.Levine,O.C.Uhlenbeck , D.N.Grothers.- J.Gralla.-Nature (New Byology), 1973, v.246, N 14, p.40-41.

4. PIPAS I., McMAHON J. Method for Predicting RNA Secondary Structure.-Proc.Nat.Acad.Sci.USA,1975,v.72,N 6,p.2017-2021.

5. БЕССОНОВ Ю.Е., СКОРОБОГАТОВ В.А. Применение относительных разбиений для поиска клик. -В кн.: Автоматизация проектирования в микроэлектронике. Теория. Методы. Алгоритмы. (Вычислительные системы, вып. 77.) Новосибирск, I978, с.24-33.

6. GAUSS D.H.,GRUTER F., SPRINZL M. Comrlication of t.N. sequences.-Nucleic Acids.Research,1979,v.6,N 1,p.r1-r64.

7. HORI H.,HIGO K.,OSAWA S. Molecular evolution of Ribosomal Components.-Proceedings of the Second Taniguchi International Symposium of Biophysics,1977. Mishima, Japan,p.240-260.

Поступила в ред.-изд.отд.
9 февраля I98I года